**Clinical Genome Analytics**

**Workflow for Whole Genome Tumor/Normal**

**Prepared by**

**Anuj Srivastava**

**Note** most of these applications will name the output based on the input. In this document, the example file names will be based on this convention. Specific input for each usage example are bracketed by < > ; parameters will be set and remain unchanged.

**Assumption**: job initiated in same directory as data files. If not, include path to data. **NOTE** that subsequent commands assume output path of prior commands will not change.

**Step1: NGS QC**

**Module Requirement:**

module load NGSQCToolkit/2.3

* 1. **Quality Statistics generation and bad read filtering**

**Illumina sequencing; CASAVA 1.8+; paired end reads; human data**

**Usage:**

IlluQC\_PRLL.pl -pe <read1.fastq> <read2.fastq> <adaptor\_file> <A> –s 30 -c 12 -o filter

**Input:**

* Forward paired end file [required]
* Reverse paired end file [required]
* Adaptor\_database file or application library [required]
* FASTQ encoding [required]

**Output:**

Folder named “filter” created here which has the following files

* read1.fastq\_filtered (here)
* read2.fastq\_filtered (here)
* \*\_unPaired\_HQReads
* \*\_stat
* \*.html,
* 6 plots for each fastq file\*\_avgQual.png, \*\_baseCompostion.png, \*\_QualRangePerBase.png, \*\_ gcDistribution.png, \*\_ qualDistribution.png, \*\_ QualRangePerBase.png,
* \*\_summary.png (summary plot)

**Parameters Description:**

* -pe: Forward reverse paired end file
* Adaptor\_file: Should be replaced by known adaptor sequence used in particular type of sequencing; if not available, use the option depending on type of sequence from below:

Primer/Adaptor libraries:

1 = Genomic DNA/Chip-seq Library

2 = Paired End DNA Library

3 = DpnII gene expression Library

4 = NlaIII gene expression Library

5 = Small RNA Library

6 = Multiplexing DNA Library

N = Do not filter for Primer/Adaptor

* FASTQ quality value variants: 5 stands for = Illumina (1.8+) (Phred+33, 33 to 7); if not known then please use A [automatic detection]; various options are given below

FASTQ variants:

1 = Sanger (Phred+33, 33 to 73)

2 = Solexa (Phred+64, 59 to 104)

3 = Illumina (1.3+) (Phred+64, 64 to 104)

4 = Illumina (1.5+) (Phred+64, 66 to 104)

5 = Illumina (1.8+) (Phred+33, 33 to 74)

A = Automatic detection of FASTQ variant

* -s: -cutOffQualScore (The cut-off value for PHRED quality score for high-quality filtering)
* -c: Number of CPUs to be used
* -o: Output will be stored in the given folder; ***provide name*** (filter here)

Note: Check here: Inside \*\_stat file that *Percentage of HQ filtered reads*” should be greater than 50 percent; if not then exit with a message that “greater than 50% of reads filtered out (data is of bad quality)

**Snippet of \*\_stat file:**

Library type Paired-end

Input files data\_1.fastq data\_2.fastq

Primer/Adaptor library Paired End DNA Library

Cut-off read length for HQ 70%

Cut-off quality score 30

Only statistics Off

Number of CPUs 12

QC statistics

File name data\_1.fastq data\_2.fastq

Total number of reads 37205349 37205349

Total number of HQ reads 33912757 33912757

Percentage of HQ reads 91.15% 91.15%

Total number of bases 2827606524 2827606524

Total number of bases in HQ reads 2577369532 2577369532

Total number of HQ bases in HQ reads 2518231490 2518391414

Percentage of HQ bases in HQ reads 97.71% 97.71%

Number of Primer/Adaptor contaminated HQ reads 3023 22

Total number of HQ filtered reads 33909713 33909713

**Percentage of HQ filtered reads 91.14% 91.14%**

Detailed QC statistics

File name data\_1.fastq data\_2.fastq data\_1.fastq\_filtered data\_2.fastq\_filtered

Minimum read length 76 76 76 76

Maximum read length 76 76 76 76

Average read length 76.00 76.00 76.00 76.00

Total number of reads 37205349 37205349 33909713 33909713

Total number of reads with non-ATGC bases 870598 309108 713977 247870

Percentage of reads with non-ATGC bases 2.34% 0.83% 2.11% 0.73%

Total number of bases 2827606524 2827606524 2577138188 2577138188

Total number of HQ bases 2669093673 2665709183 2518010286 2518167765

Percentage of HQ bases 94.39% 94.27% 97.71% 97.71%

Total number of non-ATGC bases 901356 694930 715483 321321

Percentage of non-ATGC bases 0.03% 0.02% 0.03% 0.01%

**1.2 Quality based trimming of filtered files**

**Usage:**

TrimmingReads.pl –i <filter/read1.fastq\_filtered> -irev <filter/read2.fastq\_filtered> –q 30

**Input:**

* Forward paired end file with paired end intact after filtering [required]
* Reverse paired end file with paired end intact after filtering [required]

**Output:**

* read1.fastq\_filtered\_trimmed (here)
* read2.fastq\_filtered\_trimmed (here)

**Parameters Description:**

* -i <Forward read/sequence file>
* -irev <Reverse read/sequence file of paired-end data>
* -q | -qualCutOff <Integer> (Only for FASTQ files) Cut-off PHRED quality score for trimming reads from right end (3' end)

**Reference:**

Patel et al. (2012) “NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PloS ONE*

**Step2: Alignment**

**Module Requirements:**

* module load samtools/0.1.18
* module load bwa/0.5.10

**2.1 Running BWA “aln” step**

**Usage:**

bwa aln -t 12 /pathTo/hg19.fasta <read1.fastq\_filtered\_trimmed> > aln1.sai

bwa aln -t 12 /pathTo/hg19.fasta <read2.fastq\_filtered\_trimmed> > aln2.sai

**Input:**

* Reference Genome Indexed with BWA [required]
* Read file forward; passed QC [required]
* Read file reverse; passed QC [required]

Note: Need to run bwa “aln” separately for forward and reverse file

**Output:**

* Suffix array index file; **provide file names** (aln1.sai and aln2.sai here)

**Parameters Description:**

|  |  |
| --- | --- |
| * **-t** *INT* | Number of threads (multi-threading mode) |

**2.2 Running BWA “sampe” step**

**Usage:**

bwa sampe –r @RG"\t"ID:<SAMPLE1\_RG1>"\t"LB:<SAMPLE1\_LIB1>"\t"SM:<SAMPLE1>"\t"PL:<ILLUMINA>/pathTo/reference.fasta <aln1.sai> <aln2.sai> <read1.fastq\_filtered\_trimmed> <read2.fastq\_filtered\_trimmed> > aln.sam

**Input:**

* Suffix array index file (aln1.sai and aln2.sai here) [required]
* Input fasta file (read1.fastq\_filtered\_trimmed read2.fastq\_filtered\_trimmed here)[required]
* -r: read group information (required for downstream analysis)

**Output:**

* \*.sam Alignment file in SAM format; **provide file names** (aln.sam here)

**Parameters Description:**

|  |  |
| --- | --- |
| * **-r STR** | Specify the read group in a format like ‘@RG\tID:foo\tSM:bar’. [null] |

Note: If we continue to use the same conventions as GES uses for Genologics and the HiSeq at JAX BHB, the read group info can be derived from the fastq file.

Given a pair of fastq files, the @RG info can be pulled from the name (of either of pair) and the first line of the file. Highlighted colors match read group info needed.

0916-6180\_GES12\_36721\_TTAGGC\_L007\_R2\_ALL.fastq

@HISEQ2000:158:D0R99ACXX:7:1101:1387:1965 2:N:0:TTAGGC

GTGTCTGTGGGGCCTTCCGCTCTGGCTCCTGTGGA

+

BB@FFFFFHHHHGJJFIJJIGIJJJIGJJJEBGII

@RG

ID HISEQ2000:158:D0R99ACXX:7 (the machine name, run#, flowcell ID, lane #)

LB GES12\_36721\_TTAGGC (GES sample ID with barcode)

PL ILLUMINA (platform)

SM 0916-6180 (customer sample ID)

if the given fastq file not in above format [need a script to check it]; then add timestamp to

these information:

@RG

ID ID\_timestamp

LB LIB\_timestamp

PL ILLUMINA

SM SAMPLE\_timestamp

**References:**

Li et al. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*

Li et al. (2009). "The Sequence Alignment/Map format and SAMtools." *Bioinformatics*

**Step3: Variant Pre-processing**

**Module requirements (or path):**

* module load python
* module load java
* module load R
* /opt/compsci/picard/1.8.4/
* /opt/compsci/GATK/2.2-16/GenomeAnalysisTK.jar

**Step3.1: Picard SortSam.jar (generating sorted alignment bam file)**

**Usage:**

java -Xmx2g -jar picard/SortSam.jar \

SO=coordinate \

INPUT=<aln.sam> \

OUTPUT=aln.bam \

VALIDATION\_STRINGENCY=SILENT \

CREATE\_INDEX=true

**Input:**

* The BWA generated sam file (aln.sam here) [required]

**Output:**

* The coordinate sorted alignment file in bam format; ***provide name*** (aln.bam here)
* Index of coordinate sorted alignment file (aln.bai here)

**Parameters Description:**

|  |  |
| --- | --- |
| * SO: SORT\_ORDER. If not supplied OUTPUT is in the same order as INPUT. Default value: null. Possible values: {unsorted, queryname, coordinate} |  |

* VALIDATION\_STRINGENCY: Validation stringency for all SAM files read by this program. Setting stringency to SILENT can improve performance when processing a BAM file in which variable-length data (read, qualities, tags) do not otherwise need to be decoded. Default value: STRICT. This option can be set to 'null' to clear the default value. Possible values: {STRICT, LENIENT, SILENT}
* CREATE\_INDEX: Whether to create a BAM index when writing a coordinate-sorted BAM file. Default value: false. This option can be set to 'null' to clear the default value. Possible values: {true, false}

**Step3.2: Picard Mark Duplicates (Removing duplicates in BAM file)**

**Usage:**

java -Xmx2g -jar picard/MarkDuplicates.jar \

INPUT=<aln.bam>  \

OUTPUT=aln\_dedup.bam \

METRICS\_FILE=metrics \

REMOVE\_DUPLICATES=true \

CREATE\_INDEX=true \

VALIDATION\_STRINGENCY=LENIENT

**Input:**

* Coordinated sorted BAM file (aln.bam here) [required]
* File to write duplication metrics to (metrics here) [Required]

**Output:**

* Duplicate removed bam file; ***provide name*** (aln\_dedup.bam here)
* Index of duplicate removed bam file (aln\_dedup.bai here)

**Parameters Description:**

* REMOVE\_DUPLICATES: If true do not write duplicates to the output file instead of writing them with appropriate flags set. Default value: false. This option can be set to 'null' to clear the default value. Possible values: {true, false}
* METRICS\_FILE: File to write duplication metrics

**Step3.3: Realignment around indels Part I (Target Interval Creation)**

**Usage:**

java -Xmx2g -jar GenomeAnalysisTK.jar \

-I <aln\_dedup.bam> \

-R hg19.fasta \

--num\_threads 24 \

-T RealignerTargetCreator \

-o forIndelRealigner.intervals \

-known /pathTo/Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf\

-known /pathTo/ 1000G\_phase1.indels.hg19.vcf

**Input**

* Duplicated removed bam file (aln\_dedup.bam here) [required]
* Reference genome file (hg19.fasta here) [required]

**Output**

* Interval file output to be used by IndelRealigner; ***provide name*** (forIndelRealigner.intervals here)

**Parameters Description:**

* --known: Input VCF file with known indels
* --num\_threads: How many data threads should be allocated to running this analysis

**Databases could be used with this command (-known):**

* Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf
* 1000G\_phase1.indels.hg19.vcf (currently from the 1000 Genomes Phase I indel calls)

Note: index files (\*.idx) must be available in same directory as the \*.vcf files

Note: we will use “known” files provided by the Broad GATK bundle. We will use the files that best correspond to the reference genome used.

**Step3.4: Realignment around indels Part II (Performing the local re-alignment)**

**Usage:**

java -Xmx4g -jar GenomeAnalysisTK.jar \

-I <aln\_dedup.bam> \

-R hg19.fasta \

-T IndelRealigner \

-targetIntervals <forIndelRealigner.intervals> \

-o realignedBam.bam \

-known /pathTo/Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf \

-known /pathTo/1000G\_phase1.indels.hg19.vcf

**Input:**

* Duplicated removed bam file (aln\_dedup.bam here) [required]
* Reference genome file (hg19.fasta here) [required]
* Interval file created by RealignerTargetCreator (forIndelRealigner.intervals here) [required]

**Output:**

* Bam file with realigned indels; ***provide name*** (realignedBam.bam here)
* Index file of realigned bam file (realignedBam.bai here)

**Parameters Description:**

* -known: Input VCF file with known indels

**Databases could be used with this command (-known):**

* Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf
* 1000G\_phase1.indels.hg19.vcf (currently from the 1000 Genomes Phase I indel calls)

**Step3.5: Base Quality Recalibration Part I**

**Usage:**

java -Xmx4g -jar GenomeAnalysisTK.jar \

-T BaseRecalibrator \

-I <realignedBam.bam> \

-nct 8 \

-R resources/hg19.fasta \

-knownSites /pathTo/dbsnp\_137.hg19.vcf\

-knownSites /pathTo/Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf\

-knownSites /pathTo/1000G\_phase1.indels.hg19.vcf\

-o recal\_data.grp

**Input:**

* Realigned file realignedBam.bam (here) [required]
* Reference genome file (hg19.fasta here) [required]
* -knownSites (database of known SNP and Indel sites) [not required]

**Output:**

* output recalibration table file Recal\_data.grp

**Parameters Description:**

* -knownSites: A database of known polymorphic sites to skip over in the recalibration algorithm

**Databases could be used with this command (-knownSites):**

* The most recent dbSNP release (build ID > 132) [dbsnp\_137.hg19.vcf]
* Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf
* 1000G\_phase1.indels.hg19.vcf (currently from the 1000 Genomes Phase I indel calls)

**Step3.5: Base Quality Recalibration Part II (**printing the reads**)**

**Usage:**

java -Xmx4g -jar GenomeAnalysisTK.jar \

-T PrintReads \

-R hg19.fasta \

-nct 8 \

-I <realignedBam.bam> \

-BQSR <recal\_data.grp> \

-o realigned\_BQSR.bam

**Input:**

* Realigned file realignedBam.bam (here) [required]
* Reference genome file (hg19.fasta here) [required]
* Recalibration table file (recal\_data.grp here) [required]
* Outfile name with fixed base qualities [required]

**Output:**

* Bamfile with fixed base qualities; *provide name* (realigned\_BQSR.bam here)
* Index files of Bam alignment (realigned\_BQSR.bai here)

**Parameters Description:**

* PrintReads (commands to print read from fixed BQSR file)

Note: The file size becomes almost double after BQSR step as additional information been added to sam fields.

**Step 3.6: Picard CollectAlignmentSummaryMetrics (**generating alignment statistics**)**

**Usage:**

java -Xmx2g -jar picard/CollectAlignmentSummaryMetrics.jar \ INPUT=realigned\_BQSR.bam \

OUTPUT=Metricsfile.txt \

REFERENCE\_SEQUENCE=hg19.fasta \

METRIC\_ACCUMULATION\_LEVEL=ALL\_READS \

VALIDATION\_STRINGENCY=LENIENT

**Input**

* Realigned base quality fixed file realigned\_BQSR.bam (here) [required]
* Reference sequence file [required for complete metrics]

**Output:**

* Output file with alignment statistics; ***provide name*** (Metricsfile.txt here) [required]

**References:**

* **McKenna A et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data Genome Res**
* **DePristo et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics**
* <http://picard.sourceforge.net>

**Assuming after running above steps with Tumor and Normal samples, we got two BQSR files**

**(**realigned\_BQSR\_**Normal**.bam **and** realigned\_BQSR\_**Tumor**.bam **)**

**Step4: Calling SNVs from paired tumor and normal samples**

**Module requirements (or path):**

module load somaticsniper/1.0.2

module load java

**Step4.1: SNP detection:**

**Usage:**

bam-somaticsniper –q 20 –Q 20 -F vcf -J -f hg19.fasta \

<realigned\_BQSR\_Tumor.bam> \

<realigned\_BQSR\_Normal.bam> \

<SNP.vcf>

**Note: Give the tumor and normal file in the same order**

**Input:**

* Realigned Normal base quality fixed file, realigned\_BQSR\_Normal.bam (here) [required]
* Realigned Tumor base quality fixed file, realigned\_BQSR\_Tumor.bam (here) [required]

**Output:**

* Output file detected SNPs; ***provide name*** (SNP.vcf here) [required]

**Parameter Description:**

* -f FILE REQUIRED reference sequence in the FASTA format
* -q INT filtering reads with mapping quality less than INT [0]
* -Q INT filtering somatic snv output with somatic quality less than INT [15]
* -J FLAG Use prior probabilities accounting for the somatic mutation rate
* -F STRING select output format (vcf or classic) [classic]

There are two modes, the joint genotyping mode (-J) takes into account the fact that the **tumor and normal samples are not entirely independent and also takes into account the prior probability of a somatic mutation**. This probability can be scaled to control the sensitivity of the algorithm. An accurate value for this prior would be 0.000001, but this may result in a severe lack of sensitivity at lower depths. A less realistic prior probability will generate more sensitive results at the expense of an increase in the number of false positives. To get a similar sensitivity to the default mode, we recommend using a prior of 0.01. **The default mode treats the two samples as if they came from two diﬀerent individuals**. This mode uses a less accurate mathematical model, but yields good results, especially if the normal may contain some tumor cells or the tumor is quite impure.

**Step4.2: Adding dbSNP ID:**

**Usage:**

java -Xmx2g -jar GenomeAnalysisTK.jar \

--alwaysAppendDbsnpId \

-R hg19.fasta \

-T VariantAnnotator \

-o output.vcf \

--variant <SNP.vcf> \

--dbsnp <dbsnp\_137.hg19.vcf>

**Input:**

* Somatic sniper predicted SNP file, SNP.vcf (here) [required]
* dbSNP file, dbsnp\_137.hg19.vcf (here) [required]

**Output:**

* Output file of detected SNPs; **provide name** (output.vcf here) [required]
* Output index file of detected SNPs; output.vcf.idx (here)

**Parameter Description:**

* --alwaysAppendDbsnpId: In conjunction with the dbSNP binding, append the dbSNP ID even when the variant VCF already has the ID field populated

**References:**

**Larson et al. (2011)** SomaticSniper: Identification of Somatic Point Mutations in Whole Genome Sequencing Data.***Bioinformatics***

**Step5: Calling Indels from paired Tumor and Normal samples**

**Module requirements (or path):**

* module load java
* /opt/compsci/GATK/2.2-16/GenomeAnalysisTK.jar

**Usage:**

java -Xmx2g -jar GenomeAnalysisTK.jar \

-R hg19.fasta \  
-T SomaticIndelDetector \

-o indels.vcf \

-I:normal <realigned\_BQSR\_Normal.bam> \

-I:tumor <realigned\_BQSR\_Tumor.bam> \

**Input**

* Realigned Normal base quality fixed file realigned\_BQSR\_Normal.bam (here) [required]
* Realigned Tumor base quality fixed file realigned\_BQSR\_Tumor.bam (here) [required]

**Output:**

* Output file of detected indels; ***provide name*** (indels.vcf here) [required]
* Index file of detected indels; indels.vcf.idx here

**References:**

* **McKenna A et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data Genome Res**
* **DePristo et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics**

**Broad Recommended parallelism configurations (Guide Article #1975 2013-01-14 18:02:57)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tool** | **Full name** | **Type of traversal** | **NT** | **NCT** | **SG** |
| RTC | RealignerTargetCreator | RodWalker | + | - | - |
| IR | IndelRealigner | ReadWalker | - | - | + |
| BR | BaseRecalibrator | LocusWalker | - | + | + |
| PR | PrintReads | ReadWalker | - | + | - |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tool** | **RTC** | **IR** | **BR** | **PR** |
| Available modes | NT | SG | NCT,SG | NCT |
| Cluster nodes | 1 | 4 | 4 | 1 |
| CPU threads (-nct) | 1 | 1 | 8 | 4-8 |
| Data threads (-nt) | 24 | 1 | 1 | 1 |
| Memory (Gb) | 48 | 4 | 4 | 4 |

**Broad recommended sets of known sites per tool** (http://gatkforums.broadinstitute.org/discussion/1247/what-should-i-use-as-known-variantssites-for-running-tool-x)

**Summary table**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Tool | dbSNP 129 | dbSNP >132 | Mills indels | 1KG indels | HapMap | Omni |
| RealignerTargetCreator |  |  | X | X |  |  |
| IndelRealigner |  |  | X | X |  |  |
| BaseRecalibrator |  | X | X | X |  |  |

**RealignerTargetCreator and IndelRealigner**

These tools require known indels passed with the -known argument to function properly. We use both the following files:

* Mills\_and\_1000G\_gold\_standard.indels.b37.sites.vcf
* 1000G\_phase1.indels.b37.vcf (currently from the 1000 Genomes Phase I indel calls)

**BaseRecalibrator**

This tool requires known SNPs and indels passed with the -knownSites argument to function properly. We use all the following files:

* The most recent dbSNP release (build ID > 132)
* Mills\_and\_1000G\_gold\_standard.indels.b37.sites.vcf
* 1000G\_phase1.indels.b37.vcf (currently from the 1000 Genomes Phase I indel calls)

**Appendix**

**Files required**

Human reference genome

Illumina adapter sequences used

Desired ‘known’ files from Broad GATK bundle

Interval files for exome capture bait and target (create in house from vendor’s files)